

Received June 27, 2020, accepted July 12, 2020, date of publication July 15, 2020, date of current version July 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3009541

Granular Description of Data: A Comparative Study Regarding to Different Distance Measures

FANZHONG MENG¹, CHEN FU², ZHENTANG SHI³, AND WEI LU¹², (Member, IEEE)

¹Oilfield Exploration & Production Department, SINOPEC Corp., Beijing 1100728, China

²School of Control Science and Engineering, Dalian University of Technology, Dalian 116023, China

³SINOPEC Dalian Research Institute of Petroleum and Petrochemicals, Dalian 116045, China

Corresponding author: Wei Lu (luwei@dlut.edu.cn)

This work was supported by the Fundamental Research Funds for the Chinese Central Universities under Grant DUT20LAB129.

ABSTRACT In data science, how to depict data in the concise and comprehensive way is an important issue. To address the issue, the key is to construct descriptors that are highly interpretable and can be used to reveal the data structure. Information granules, as one important role in the field of granular computing, are entities that can be easily represented and abstracted from data. Therefore, by constructing a series of information granules, the characteristics of data can be captured and described, and the granular description of data is realized. A key part of the granular description of data is to explore the geometric characteristics (locations and shapes) of information granules used to describe data. Since distance measures directly affect the geometric characteristics of the constructed information granules, a comparative study based on three different distance measures is conducted in this paper. From the experimental results based on both synthetic and UCI repository datasets, it can be seen that the information granules constructed in the case where three different distance measures are used show different geometrical shapes, and can describe the data in a concise way. Furthermore, the data structure can be explored more comprehensively by using three distance measures.

INDEX TERMS Data description, granular computing, hypersphere information granules, distance measures.


I. INTRODUCTORY COMMENTS

Data description [1], [2], which can reveal the nature of data, is playing an increasingly pivotal role in the field of data analysis. The key objective of data description is to construct both highly interpretable and concise descriptors which help better reveal the structure of original data or construct classification models. The problem of the objective is exacerbated for the diversity of data structure, viz., data with high dimensionality and data with special geometric structure. Therefore, it's necessary to find an appropriate method to abstract and extract knowledge from data and information to construct descriptors for different datasets.

Granular computing (GrC) [3], as a new and rapidly developing information processing paradigm, consists of a series of concepts, methods and applications. Yao *et al.* [4] reviewed the foundations of GrC and elaborated on the development of its research. GrC can be regarded as a human-centered representation and processing of knowledge [5] and used to help adjust the levels of abstraction of data more flexibly [6]. Therefore, GrC has some practical applications such as the

resilience analysis of critical infrastructures [7] and fuzzy time series forecasting [8]. Information granules [9], which is the major research object of GrC, are a series of data entities that are to some extent (granularity) [10] indivisible due to the content similarity or functional proximity. They are regarded as the abstraction of the perceived concepts which can be expressed by some existing formalisms such as intervals [11], fuzzy sets [12], [13], rough sets [14] and shadowed sets [15] among others. The process of constructing information granule is a process of understanding the structure, distribution and content of data, which helps us explore and extract knowledge from data more intuitively. For instance, while encountering a set of planar data with two different regions, see Fig. 1. Intuitively, we can describe the two regions by forming three information granules to capture their topological geometric structures.

As a result, granular description of data with information granules has been witnessed in a number of pursuits of data analytics. Pedrycz *et al.* [2] presented a two-stage framework in which triangular information granules are designed to describe the nature of numeric data. Some numeric prototypes are first obtained by fuzzy C-means clustering (FCM). For each attribute of the data, the principle of justifiable

The associate editor coordinating the review of this manuscript and approving it for publication was Hong-Mei Zhang .

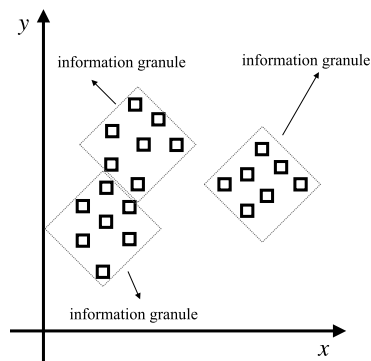


FIGURE 1. The topological geometric structures of two regions of data are described by forming three information granules.

granularity is used to construct information granules around the corresponding attribute of the prototypes. The information granules are constructed according to the coverage and specificity on each dimension of data in the form of triangular information granule and the characteristics of the dominant structure of data are captured. Zhu *et al.* [16] also use a similar two-stage framework to build a granular descriptor with ellipsoidal information granules. After using clustering methods to generate the numeric prototypes, the ellipsoidal information granules are formed around the prototypes and the lengths of their semi-axes are optimized. Spherical information granules [17] and hypercube information granules [18] are also formed in similar manners for data description. In the research proposed by Lu *et al.* [19], a series of information granules with multiple values of granularity are constructed through engaging a synergy between cone-shaped fuzzy sets and the concept of information granularity, so as to realize the granular description of multidimensional data. For the sake of depicting the geometric shapes of data distribution more accurately, Ouyang *et al.* [20] continuously used twice clustering (DBSCAN and FCM) to obtain numeric prototypes with more reasonable locations. Then, around these prototypes and corresponding partition data, the principle of justifiable granularity is adopted to build corresponding information granules. Similarly, Fu *et al.* [21] applied the idea of granular description to data classification by constructing a collection of information granules for each class of data, and then unite the information granules to capture the key characteristics of each class of data.

The above-mentioned researches have contributed to the granular description of data. In particular, this paper, together with the studies in [19] and [21], realizes the granular description of data according to the fundamental framework proposed by Pedrycz *et al.* [2]. However, these two studies and other studies use only one specific kind of distance measure to produce information granules, and the resulted information granules have just one certain geometric shape (one of rectangular, ellipsoid, sphere, hyper-box and hyper-cube, etc.), which affects the performance of the resulting granular descriptors to describe the data. In other words, while encountering some datasets with unique geometric distributions, the information granules produced with a single

type of distance measure lack the ability to analyze and reveal the details of data from multiple perspectives. Unlike the “multiple information granularity” in [19] and the “union information granule” in [21], the ultimate objective of this article is focusing on using three different distance measures to generate hypersphere information granules with three different geometric shapes (sphere, cube and diamond) to carry out a comparative study instead of focusing on changing the sizes or quantities of information granules. By accomplishing this objective, we can perform granular description of data from three perspectives compared to the methods proposed by the existing papers. At the first stage of the proposed granular description method, data are partitioned into some data chunks by means of FCM clustering [22], [23]. And then, the corresponding information granules are constructed on the data chunks they belong to at the following second stage. The last stage focuses on further refining the information granules by eliminating the overlaps between them. Through exploiting three distance measures, viz., Euclidean, Chebyshev, and Manhattan distances, to carry out the above three stages, three collections of hypersphere information granules with different geometric shapes are formed in turn to describe the data. Compared to the data description studies mentioned above, this paper exhibits the following original aspects:

- Hypersphere information granules with three different geometric shapes are generated by using three different distance measures, respectively.
- The data can be described from multiple perspectives by constructing hypersphere information granules with different geometric shapes.
- The proposed granular description method produces hypersphere information granules to describe data with simple architecture and reliable performance.

The paper is structured as follows. We present the representation of hypersphere information granules regarding three different distance measures, and introduce the criteria of coverage and specificity of information granules in Section II. The approach of granular description of data is introduced in detail in Section III. In Section IV, some experiments on synthetic and publicly available (UCI) datasets are completed to visualize and analyze the feasibility of the granular description approach. Besides, the impact of using three different distance measures on the granular description is also analyzed. Section V concludes the paper.

II. REPRESENTATION OF HYPERSPHERE INFORMATION GRANULES AND ITS COVERAGE AND SPECIFICITY

Since our proposed granular description of data is a comparative study in which hypersphere information granules are constructed using three distance measures, in this section, we first explain the representation of a hypersphere information granule regarding to three different distance measures, viz., Euclidean distance, Chebyshev distance, and Manhattan distance, respectively. Next, based on the framework of constructing one-dimensional interval information

granules proposed by Pedrycz in [10], we introduce the criteria of coverage and specificity as the quantitative standard for constructing multidimensional hypersphere information granules.

A. REPRESENTATION OF HYPERSPHERE INFORMATION GRANULES

While faced with cognitive and decision-making activities, human beings tend to put information and data together due to their similarity in advance. This process of organizing and abstracting knowledge from information and data is called information granulation which results in a series of meaningful entities, i.e., information granules. Apparently, numerical data or information can be abstracted into information granules by means of information granulation. For example, fuzzy clustering methods and the principle of justifiable granularity can be adopted to produce information granules and then realize the granular description of data.

Through abstract methods (information granulation), a collection of normalized n -dimensional data, viz., $\mathbf{X} = \{\mathbf{x}_t \in [0, 1]^n | t = 1, 2, \dots, N\}$, can be represented by an information granule expressed as follows,

$$\Omega = \{\mathbf{x}_k \mid \|\mathbf{x}_k - \mathbf{v}\|_d \leq \rho, \mathbf{x}_k \in \mathbf{X}\}, \quad (1)$$

where \mathbf{v} expresses the representative (center) of the information granule and $\|\mathbf{x}_k - \mathbf{v}\|_d$ expresses the Minkowski distance between point \mathbf{x}_k and center \mathbf{v} . Obviously, the information granule Ω covers the samples coming from \mathbf{X} whose distance versus \mathbf{v} is less than a predefined radius ρ . All points in space $[0, 1]^n$ with a distance equalling ρ from the center \mathbf{v} constitute the geometric surface of the information granule Ω .

For two points in $[0, 1]^n$, i.e., $\mathbf{x}_a = (x_{a1}, x_{a2}, \dots, x_{an})$ and $\mathbf{x}_b = (x_{b1}, x_{b2}, \dots, x_{bn})$, the Minkowski distance is calculated by

$$\|\mathbf{x}_k - \mathbf{v}\|_d = (|x_{a1} - x_{b1}|^p + \dots + |x_{an} - x_{bn}|^p)^{\frac{1}{p}}. \quad (2)$$

When the parameter p is set as 2, we obtain the Euclidean distance (denoted as “ $\|\cdot\|_E$ ”), which denotes the “ordinary” straight-line distance between \mathbf{x}_a and \mathbf{x}_b , viz.,

$$\|\mathbf{x}_a - \mathbf{x}_b\|_E = \sqrt{(x_{a1} - x_{b1})^2 + \dots + (x_{an} - x_{bn})^2}.$$

When the parameter p goes to infinity, we obtain the Chebyshev distance (denoted as “ $\|\cdot\|_C$ ”), which denotes the greatest of the absolute differences along any coordinate dimension between \mathbf{x}_a and \mathbf{x}_b , viz.,

$$\|\mathbf{x}_a - \mathbf{x}_b\|_C = \max\{|x_{a1} - x_{b1}|, \dots, |x_{an} - x_{bn}|\}.$$

When the parameter p is set as 1, we obtain the Manhattan distance (denoted as “ $\|\cdot\|_M$ ”), which denotes the sum of the absolute differences of the Cartesian coordinates between \mathbf{x}_a and \mathbf{x}_b , viz.,

$$\|\mathbf{x}_a - \mathbf{x}_b\|_M = |x_{a1} - x_{b1}| + \dots + |x_{an} - x_{bn}|.$$

Fig. 2 presents an example of the three different distances between two-dimensional points, where the length of the red line stands for the Euclidean distance, the length of the blue

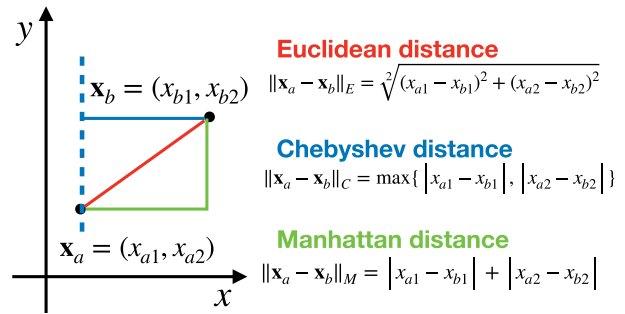


FIGURE 2. Three distance measures for two-dimensional data.

solid line stands for the Chebyshev distance and the length of the green line stands for the Manhattan distance.

Choosing different parameter p of the Minkowski distance will directly affect the geometry of the corresponding information granules. When the parameter p is set as 2, ∞ and 1, respectively, the Euclidean, Chebyshev and Manhattan distances are used to construct information granules with regular geometric shapes, see Fig. 3. Apparently, if the dimension of data, i.e., n , is 2, the information granule Ω will emerge in the form of circle, square, and diamond, respectively. Whereas if n is 3, the information granule Ω is presented in the form of a sphere, cube and octahedron, respectively. Also when the size of the information granule Ω , viz., ρ equals zero, Ω degenerated as a point \mathbf{v} (i.e., the center of Ω). It is noteworthy that when the parameter p in (2) is set as other values, the resulting information granules appear as the irregular geometries, which means that these resulting information granules with the irregular geometries can not be represented in the way of simple easy operation such that they are rarely used in granular description. Therefore, in this study, Euclidean, Chebyshev and Manhattan distances obtained by respectively setting the parameter p as 2, ∞ and 1 are used to construct information granules with different geometry which can be easily represented by (1) with the symbols \mathbf{v} and ρ expressing the center and radius. We collectively call these resulting information granules which exhibit specific geometric shapes with regard to three different distance measures hypersphere information granules.

B. THE COVERAGE AND SPECIFICITY OF A HYPERSPHERE INFORMATION GRANULE

When constructing an information granule around a cluster of data with a prototype, the center of the information granule to be built is directly determined with the prototype of data. We only need to focus on how to determine another parameter, viz., radius. The rationale behind this is that we should find a quantitative standard when constructing information granules, so that we can make a clear and concise description for data where (i) the constructed information granule is *justified* based on the experimental data, (ii) the *semantics* of the produced information granule is well-defined which means that we can easily separate it from other ones. We can quantify these two requirements with the aid of the criteria of coverage and specificity, respectively.

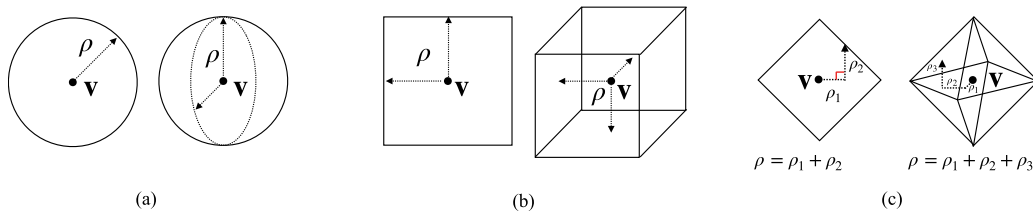


FIGURE 3. Geometry of information granules induced by the (a) Euclidean, (b) Chebyshev and (c) Manhattan distances.

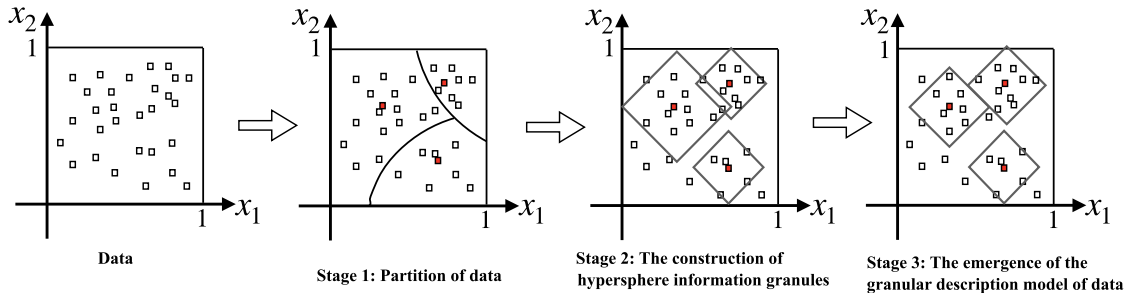


FIGURE 4. The blueprint of the proposed method of granular description using Manhattan distance as an example.

The coverage of an information granule is usually used to quantify the rationality of the relationship between information granules and the corresponding data to be described. Just as its name stipulates, the coverage of an information granule reveals the extent to which information granule covers data to be described. The more data covered by an information granule, the higher the coverage of information granule is and the more rational the granule is. The specificity of an information granule reflects a degree to which the information granule describes the details of experimental data. This property is closely related to the measure of an information granule. The smaller an information granule is, the more specific it is, and more meaningful the information granule turns to be. For an information granule with only one single element, i.e., the size is zero, we assume that the information granule owns the highest specificity. Whereas if all experimental data is covered by an information granule, we assume that it owns the lowest specificity.

Generally, the coverage can be quantified according to the quantity of elements that are covered by an information granule. Consider the normalized dataset $\mathbf{X} = \{\mathbf{x}_i \mid \mathbf{x}_i \in [0, 1]^n, i = 1, 2, \dots, N\}$, the coverage of the information granule to be constructed on \mathbf{X} can be defined as follows:

$$\text{Cov}(\Omega) = \text{card}\{\mathbf{x}_k \in \Omega, \mathbf{x}_k \in \mathbf{X}\}, \quad (3)$$

where $\Omega = \{\mathbf{x}_k \mid \|\mathbf{x}_k - \mathbf{v}\|_d \leq \rho, \mathbf{x}_k \in \mathbf{X}\}$ is a hypersphere information granule defined on \mathbf{X} and $\text{card}\{\cdot\}$ stands for the count of elements in \mathbf{X} falling within the border of Ω . The specificity of the information granule to be constructed on \mathbf{X} can be quantified by a non-increasing function of the size of corresponding information granule, viz.,

$$\text{Spec}(\Omega) = 1 - \rho. \quad (4)$$

We can see clearly that the specificity of the hypersphere information granule Ω is only related to one variable, i.e., the radius ρ .

Apparently, the quantifications of coverage and specificity contradict each other: the growth of values of one results in the decrease of the other one. Therefore, there should be a balance between the two properties, which can be determined by the center \mathbf{v} and radius ρ to maximize the product of them, says,

$$\arg \max_{\mathbf{v}, \rho} \{\text{Cov}(\Omega) \times \text{Spec}(\Omega)\}. \quad (5)$$

The symbol ‘‘arg max’’ means that what we need to calculate is the values of \mathbf{v} and ρ which make the value of $\text{Cov}(\Omega) \times \text{Spec}(\Omega)$ reaching its maximum.

III. THE PROPOSED METHOD OF GRANULAR DESCRIPTION OF DATA WITH THREE DIFFERENT DISTANCE MEASURES

In this section, the proposed strategy of granular description of data is presented. A normalized n -dimensional dataset \mathbf{D} is considered, where $\mathbf{D} = \{\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{in}) \mid \mathbf{x}_i \in [0, 1]^n, i = 1, 2, \dots, N\}$, of which, N denotes the count of elements in \mathbf{D} . The structure of the proposed granular description strategy is shown in Fig. 4.

A. PARTITION OF DATASET

The task of this stage is to partition the dataset \mathbf{D} through invoking a clustering algorithm, viz., FCM. A series of prototypes are first generated by invoking FCM on \mathbf{D} , which are treated as the centers of the constructed hypersphere information granules. The cluster number equals c , where $c \geq 2$. With the accomplishment of clustering, c prototypes, that is, $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_c$, and a c by N partition matrix $\mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_N] = [u_{ji}]_{\substack{j=1,2,\dots,c \\ i=1,2,\dots,N}}$ are returned. More intuitively, partition matrix \mathbf{U} is expressed as follows,

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1N} \\ u_{21} & u_{22} & \dots & u_{2N} \\ \vdots & \vdots & u_{ji} & \vdots \\ u_{c1} & u_{c2} & \dots & u_{cN} \end{bmatrix}. \quad (6)$$

The entry u_{ji} standing for the membership degree of the i th sample \mathbf{x}_i in the dataset \mathbf{D} versus the prototype \mathbf{v}_j is determined by the formula

$$u_{ji} = \frac{1}{\sum_{e=1}^c \left(\frac{\|\mathbf{x}_i - \mathbf{v}_j\|_d}{\|\mathbf{x}_i - \mathbf{v}_e\|_d} \right)^{\frac{2}{m-1}}}, \quad j = 1, 2, \dots, c, \quad (7)$$

where m is a fuzzification coefficient, whose value is usually set as 2. The prototypes are computed as follows:

$$\mathbf{v}_j = \frac{\sum_{i=1}^N u_{ji}^m \mathbf{x}_i}{\sum_{i=1}^N u_{ji}^m}, \quad j = 1, 2, \dots, c \quad (8)$$

Also, it is worth noting that a distance formula $\|\mathbf{x}_i - \mathbf{v}_j\|_d$ is used here. Considering that the paper is a comparative study, the distance $\|\mathbf{x}_i - \mathbf{v}_j\|_d$ should be calculated by using three distance measures, viz., Euclidean, Chebyshev and Manhattan distances. Subsequently, the i th sample \mathbf{x}_i in \mathbf{D} is attached to the s th prototype \mathbf{v}_s by choosing the highest membership degree from the k th attribute in \mathbf{U} , i.e., $s = \arg \max_{j=1,2,\dots,c} \{u_{ji}\}$. In this way, all samples attached to the s th prototype \mathbf{v}_s are grouped and we obtain a chunk for \mathbf{v}_s , i.e., \mathbf{D}_s . As a result, the dataset \mathbf{D} is divided into c chunks $\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_c$ by traversing all elements in dataset \mathbf{D} , where $\mathbf{D}_j = \{\mathbf{x}_{ij} = (x_{1j}, x_{2j}, \dots, x_{N_{ij}}) \mid \mathbf{x}_{ij} \in \mathbf{D}, i_j = 1, 2, \dots, N_j\}$ with $j = 1, 2, \dots, c$. In the following stage, the hypersphere information granules are constructed on these produced chunks.

B. THE CONSTRUCTION OF HYPERSPHERE INFORMATION GRANULES ON INDIVIDUAL CHUNKS

This stage focuses on building initial hypersphere information granules around the chunks. The hypersphere information granules include two key parameters, i.e., their centers and radii. Focusing on these two parameters, the hypersphere information granules can be constructed.

Considering the j th chunk \mathbf{D}_j , the hypersphere information granule Ω_j comes with the form (1) is formed. To determine its center and radius, two requirements are encountered: (i) the hypersphere information granule to be formed should cover as more elements from \mathbf{D}_j as possible, and (ii) the length of its radius should be as shorter as possible. With regard to the concepts introduced in Subsection II-B, these two parameters can be fixed by measuring the coverage and specificity of the hypersphere information granule Ω_j . More specifically, the coverage and specificity can be defined as

$$\text{Cov}(\Omega_j) = \text{card}\{\mathbf{x}_i \mid \|\mathbf{x}_i - \mathbf{v}_j\|_d \leq \rho_j, \mathbf{x}_i \in \mathbf{D}_j\} \quad (9)$$

and

$$\text{Spec}(\Omega_j) = 1 - \rho_j, \quad (10)$$

respectively, where \mathbf{v}_j as the prototype produced on chunk \mathbf{D}_j is set as the center of Ω_j , ρ_j is treated as the radius to be optimized. Since the center has been defined as \mathbf{v}_j ,

according to (5), the coverage and specificity can be balanced by only changing the value of radius ρ_j which means that the radius ρ_j is the only argument. By traversing all the samples in chunk \mathbf{D}_j , we calculate their distances to the center \mathbf{v}_j , and let these N_j distances, viz., $d_{ij} = \|\mathbf{x}_{ij} - \mathbf{v}_j\|_d$ with $i_j = 1, 2, \dots, N_j$, be the radius of the hypersphere information granule, respectively. Then, the optimized radius of Ω_j can be calculated by

$$\rho_j^{opt} = \arg \max_{\rho_j=d_{1j}, d_{2j}, \dots, d_{N_{ij}}} \{\text{Cov}(\Omega_j) \times \text{Spec}(\Omega_j)\}. \quad (11)$$

Further, for all c chunks $\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_c$, the corresponding hypersphere information granules can be constructed in the same way. So far, c initial hypersphere information granules $\Omega_1, \Omega_2, \dots, \Omega_c$ are constructed on \mathbf{D} , where $\Omega_j = \{\mathbf{x}_{ij} \mid \|\mathbf{x}_{ij} - \mathbf{v}_j\|_d \leq \rho_j^{opt}, \mathbf{x}_{ij} \in \mathbf{D}_j\}$ with $i_j = 1, 2, \dots, N_j$ and $j = 1, 2, \dots, c$.

C. THE EMERGENCE AND EVALUATION OF THE HYPERSPHERE INFORMATION GRANULE-BASED GRANULAR DESCRIPTOR

Note that information granules inherently carry with semantics, which means that c initial hypersphere information granules $\Omega_1, \Omega_2, \dots, \Omega_c$ formed on the chunks $\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_c$ should be distinguished and non-overlapping from each other. Therefore, the aim is to make these c hypersphere information granules do not overlap with each other and then to form the granular descriptors of data with the refined hypersphere information granules.

For any two hypersphere information granules $\Omega_a = \{\mathbf{x}_{ia} \mid \|\mathbf{x}_{ia} - \mathbf{v}_a\|_d \leq \rho_a^{opt}, \mathbf{x}_{ia} \in \mathbf{D}_a\}$ and $\Omega_b = \{\mathbf{x}_{ib} \mid \|\mathbf{x}_{ib} - \mathbf{v}_b\|_d \leq \rho_b^{opt}, \mathbf{x}_{ib} \in \mathbf{D}_b\}$ where $a \neq b, a, b = 1, 2, \dots, c$. If the distance between their centers, i.e., $d_{ab} = \|\mathbf{v}_a - \mathbf{v}_b\|_d$, is less than the sum of their radius ρ_a^{opt} and ρ_b^{opt} , these two hypersphere information granules overlap with each other. In order to dismiss the overlaps, we scale the radii of the hyperspheres Ω_a and Ω_b to the half of the distance between their centers, by making the compromise between the coverage and specificity of these two hyperspheres, respectively, viz.,

$$\rho_a^{adj} = \rho_b^{adj} = \frac{\|\mathbf{v}_a - \mathbf{v}_b\|_d}{2}, \quad (12)$$

which results in two non-overlapping two hypersphere information granules $\Omega_a^{adj} = \{\mathbf{x}_{ia} \mid \|\mathbf{x}_{ia} - \mathbf{v}_a\|_d \leq \rho_a^{adj}, \mathbf{x}_{ia} \in \mathbf{D}_a\}$ and $\Omega_b^{adj} = \{\mathbf{x}_{ib} \mid \|\mathbf{x}_{ib} - \mathbf{v}_b\|_d \leq \rho_b^{adj}, \mathbf{x}_{ib} \in \mathbf{D}_b\}$.

Refer to Fig. 5, the eliminations of overlap for hypersphere information granules induced by the Euclidean, Chebyshev and Manhattan distances are shown separately. It should be noted that no matter which distance measure is used here, it must be consistent with the measures used in the previous two stages. Through traversing any two hypersphere information granules from $\Omega_1, \Omega_2, \dots, \Omega_c$, the granular descriptor constructed by c refined hypersphere information granules $\Omega_1^{adj}, \Omega_2^{adj}, \dots, \Omega_c^{adj}$ are completely produced.

To evaluate the description of data, we focus on the representation capabilities of the produced granular descriptors,

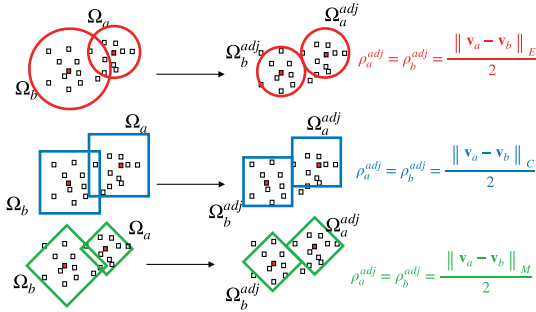


FIGURE 5. The process of elimination of overlaps among all c produced hypersphere information granules.

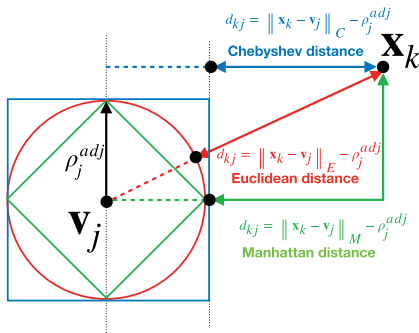


FIGURE 6. The calculation of distances between \mathbf{x}_k and three hypersphere information granules with different geometric structures.

and the ability to reconstruct numeric data for the hypersphere information granules is considered. Given a numeric element $\mathbf{x}_k \in [0, 1]^n$ inputs the granular descriptor, its membership degrees versus individual hypersphere information granules can be obtained by calculating the distance between the element \mathbf{x}_k and the hypersphere information granule Ω_j^{adj} ($j = 1, 2, \dots, c$). Two scenarios should be discussed while calculating the distance, i.e.,

- 1) $\mathbf{x}_k \in \Omega_j^{adj}$: In this scenario, the element \mathbf{x}_k falls inside the border of hypersphere information granule Ω_j^{adj} . Apparently, the distance between \mathbf{x}_k and Ω_j^{adj} , viz., d_{kj} , equals zero, i.e., $d_{kj} = 0$.
- 2) $\mathbf{x}_k \notin \Omega_j^{adj}$: In this scenario, the element \mathbf{x}_k does not fall inside the border of any hypersphere information granule Ω_j^{adj} . Referring to Fig. 6, the distance d_{kj} can be measured by finding out the shortest distance between the element and the granule Ω_j^{adj} , i.e.,

$$d_{kj} = \begin{cases} \|\mathbf{x}_k - \mathbf{v}_j\|_E - \rho_j^{adj}, & \text{Euclidean} \\ \|\mathbf{x}_k - \mathbf{v}_j\|_C - \rho_j^{adj}, & \text{Chebyshev} \\ \|\mathbf{x}_k - \mathbf{v}_j\|_M - \rho_j^{adj}, & \text{Manhattan.} \end{cases} \quad (13)$$

Having the distance between the element \mathbf{x}_k and the j th granule Ω_j^{adj} , the membership degree μ_{kj} of \mathbf{x}_k versus the Ω_j^{adj} can be calculated through the following expression

$$\mu_{kj} = \begin{cases} 1, & \text{if } \mathbf{x}_k \in \Omega_j^{adj} \\ \frac{1}{\sum_{e=1}^c \left(\frac{d_{kj}}{d_{ke}}\right)^{\frac{2}{m-1}}}, & \text{if } \mathbf{x}_k \notin \Omega_j^{adj}. \end{cases} \quad (14)$$

Further, in light of the membership degree, a granular result of \mathbf{x}_k can be reconstructed by,

$$\hat{\mathbf{v}}_k = \frac{\sum_{j=1}^c \mu_{kj}^m \mathbf{v}_j}{\sum_{j=1}^c \mu_{kj}^m} \quad (15)$$

$$\hat{\rho}_k = \frac{\sum_{j=1}^c \mu_{kj}^m \rho_j^{adj}}{\sum_{j=1}^c \mu_{kj}^m}, \quad (16)$$

and the quality of reconstruction uses the coverage criterion formed as denoted follows

$$Q_{Acc} = \sum_{k=1}^N \frac{T(\mathbf{x}_k)}{N} \times 100,$$

$$\text{with } T(\mathbf{x}_k) = \begin{cases} 1, & \text{if } \|\mathbf{x}_k - \hat{\mathbf{v}}_k\|_d \leq \hat{\rho}_k \\ 0, & \text{otherwise} \end{cases}. \quad (17)$$

IV. EXPERIMENTAL STUDIES

Two synthetic datasets and four publicly available datasets from the UCI repository are considered in the related experiments in this section. There are two objectives of the experiments are (i) to visualize and validate the feasibility of the proposed approach of granular description and (ii) to explore the impact of the different distance measures used in the proposed method on the performance of the granular description of data.

Before all experiments, we preprocess all the datasets by normalizing the values on every attribute of them into a unit interval. For the sake of carrying out a comparative study, all distance calculations involved in the process presented in Section III are completed by using Euclidean, Chebyshev, and Manhattan distances, respectively in the ensuing experiments, and then three groups of hypersphere information granules with different geometric structures are produced. Therefore, for each dataset, the corresponding granular description is evaluated by Q_{Acc} which is calculated by (17). Besides, the number of clusters c ranges from 2 to 10 with step 1.

A. SYNTHETIC DATASETS

There are two synthetic datasets showing different geometric structures, which are generated in the following way.

- 1) Blobs dataset, which consists of three groups of data, includes 600 samples totally. The i th ($i = 1, 2, 3$) group with 200 samples is generated according to the Normal distribution with the mean vector μ_i and the covariance matrix Σ_i , where $\mu_1 = [4, 2]$, $\Sigma_1 = \begin{bmatrix} 0.8 & 0 \\ 0 & 0.3 \end{bmatrix}$, $\mu_2 = [1, 7]$, $\Sigma_2 = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.5 \end{bmatrix}$, $\mu_3 = [5, 6]$, $\Sigma_3 = \begin{bmatrix} 1.1 & 0 \\ 0 & 1.7 \end{bmatrix}$.
- 2) Square dataset, which consists of two square shape groups of data, includes 800 instances totally. The i th ($i = 1, 2$) group is generated by uniformly distributing 400 samples on a square with a side length of 1

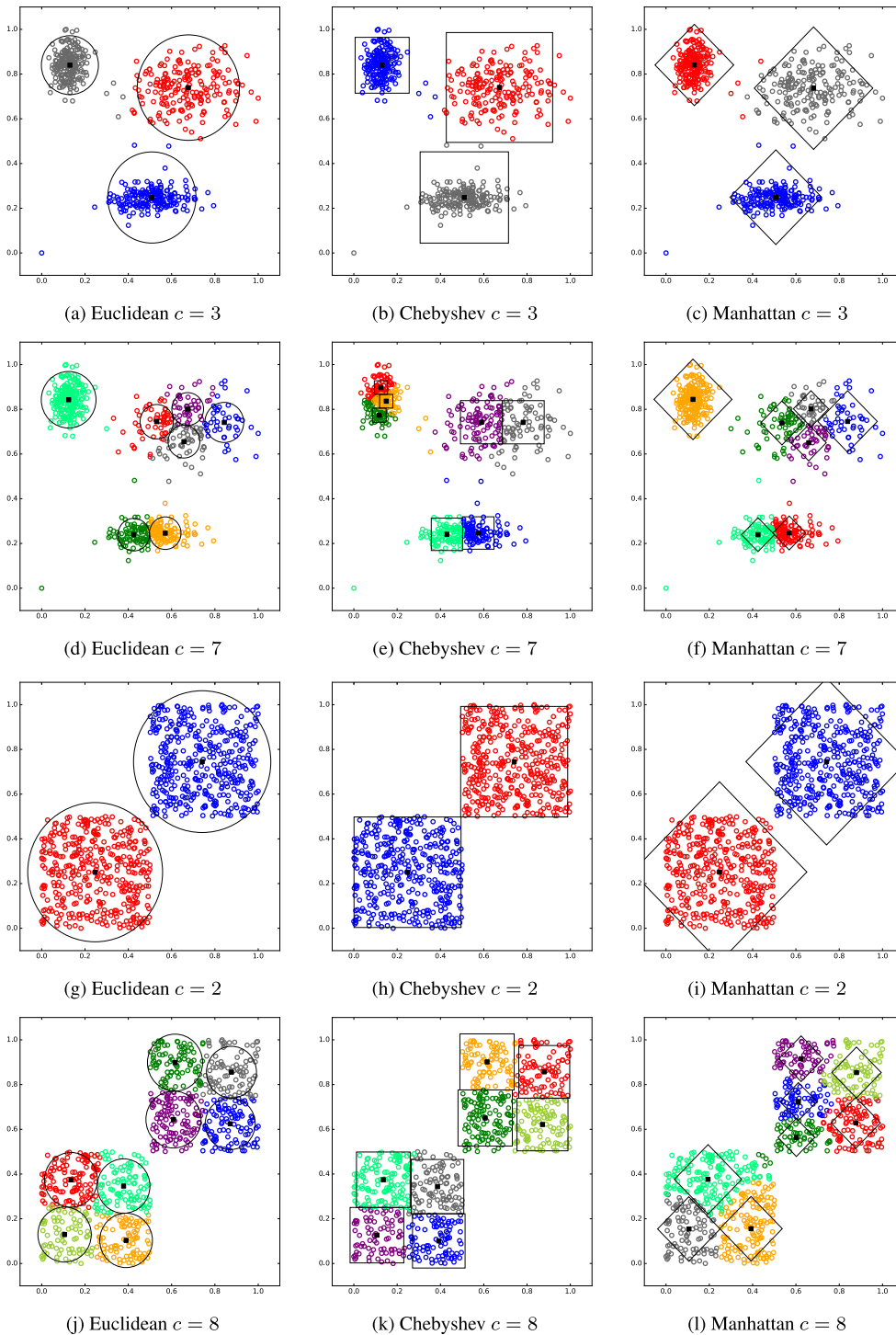


FIGURE 7. The hypersphere information granules formed based on three distance measures for the synthetic datasets.

whose center located at \mathbf{x}_i , where $\mathbf{x}_1 = (0.5, 0.5)$ and $\mathbf{x}_2 = (1.5, 1.5)$.

We experimented on the above two synthetic datasets and the reconstruction quality Q_{Acc} as the experimental results are reported in Fig. 9.

For Blobs dataset (see Fig. 9a), the value of Q_{Acc} reaches its maximum at 95.67%, 96.34% and 92.51% for all three distance measures when the value of c is 3. In this case

($c = 3$), the established hypersphere information granules for describing Blobs dataset are shown in Fig.7a, 7b and 7c. We can see clearly that hypersphere information granules with all three different shapes, i.e., circles, squares, and diamonds, can well describe the Blobs dataset. When c is set as 7, the value of Q_{Acc} obtained by using Chebyshev distance (see the blue line in Fig. 9a) encounters a sharp downtrend. In contrast, the values of Q_{Acc} obtained

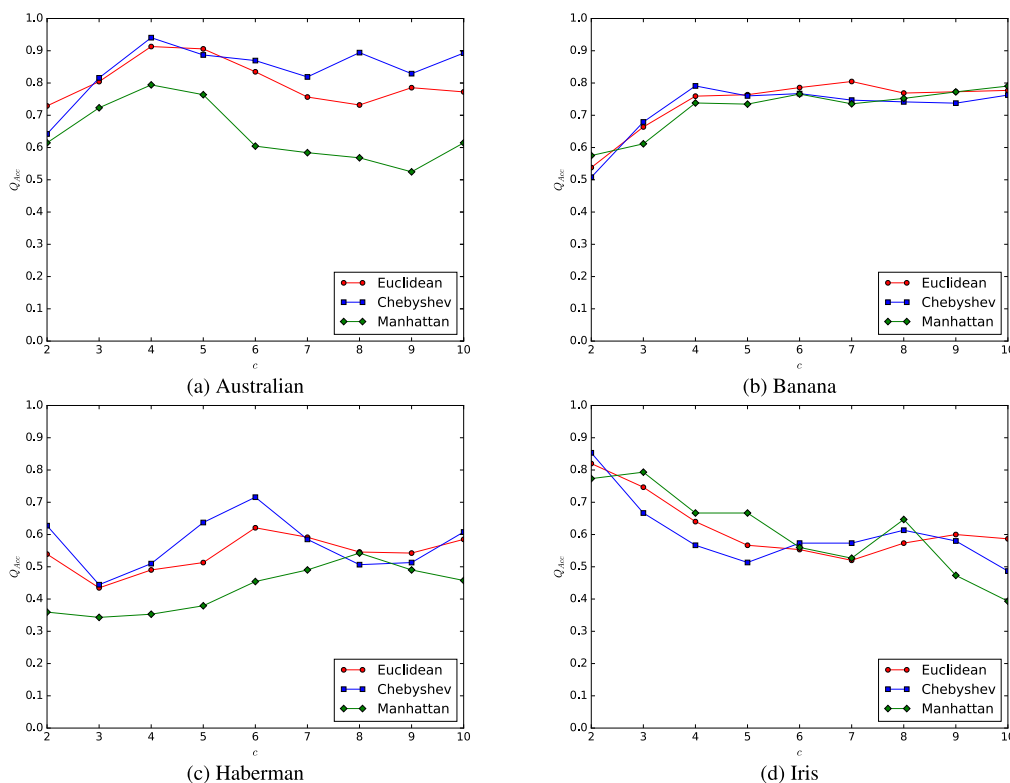


FIGURE 8. Experimental results of the publicly available datasets using different distance measures versus the values of the cluster number c .

by using Euclidean and Manhattan distances (see the red and green lines in Fig.9a) exhibit an upward trend. Let us focus on the cluster located at the top left corner of Blobs dataset, the hypersphere information granules established by Euclidean (one circle in Fig.7d) and Manhattan (one diamond Fig. 7f) distances are indeed more capable of capturing the geometric structure than the ones (three squares in Fig. 7e) built by using Chebyshev distance.

Unlike the case that the best description performances of three distance measures all appear at $c = 3$ for Blobs dataset, the best description performances of the hypersphere information granules established by three distance measures on Square dataset appear at different c values. According to the red line presented in Fig. 9b, the value of Q_{Acc} reaches its maximum at 94.13% with $c = 2$ when using Euclidean distance to construct hypersphere information granules. The hypersphere information granules in the shape of circles in Fig. 7g strike a better balance between precise coverage and justifiable separation of two squares data than the ones in the shape of squares (Fig. 7h) and diamonds (Fig. 7i). However, a “higher” maximum 94.75% with $c = 8$ is encountered (see the blue line presented in Fig. 9b) when using Chebyshev distance, the related information granules is shown in Fig. 7k. We can see clearly that the hypersphere information granules formed by using Chebyshev distance with $c = 8$ own geometric shapes that fit well with the Square dataset.

Based on the experimental results of the above two synthetic datasets, it can be found that datasets with different geometric structures have the following relationship with

TABLE 1. Summary of publicly available datasets involved in the experiments.

Dataset	Samples	Attributes	Classes
Australian	690	14	2
Banana	5300	2	2
Haberman	326	3	2
Iris	150	4	3

the used distance measures when constructing hypersphere information granules. For a dataset consisting of several separated clusters without sharp geometric structures (such as Blobs dataset), the description quality of the hypersphere information granules built by using Euclidean and Chebyshev distances is slightly better than the ones built by using Manhattan distance. The main reason behind this is that when describing these datasets, the hypersphere information granules constructed by using the first two distances can cover as many data points as possible without containing much blank space. For the datasets with sharp geometric shapes (such as Square datasets), the information granules with corresponding shapes (such as the square-shape information granules generated by Chebyshev distance) can describe the datasets more accurately than the ones generated by other two distance measures.

B. PUBLICLY AVAILABLE DATASETS

Four publicly available datasets coming from the UCI repository (<http://archive.ics.uci.edu/ml/index.php>) are also considered in experiments. These datasets are summarized in Tab. 1, showing the dataset name, the number

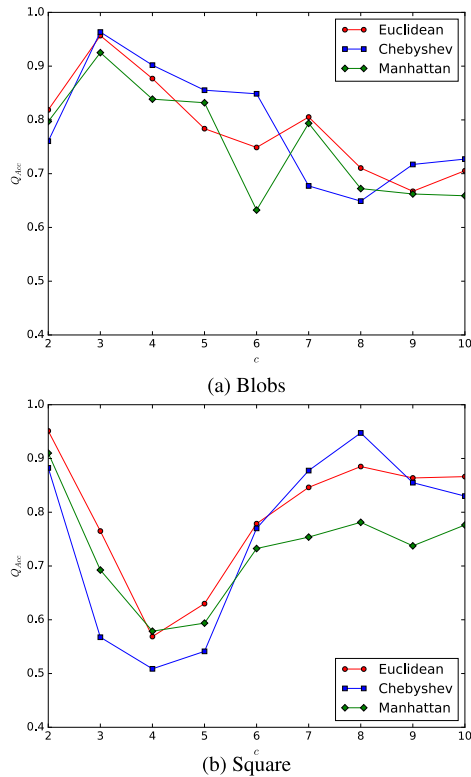


FIGURE 9. Experimental results of the synthetic datasets using different distance measures versus the values of the cluster number c .

of samples, attributes and classes. The experimental results for individual datasets are reported in Fig. 8.

From Fig. 8a and 8c, we can see that the description performances of the information granules constructed by using Chebyshev distance show the best for Australian dataset ($Q_{Acc} = 94.06\%$ with $c = 4$) and Haberman dataset ($Q_{Acc} = 71.57\%$ with $c = 6$). For Banana dataset, see the changing trend of values of Q_{Acc} with c in Fig. 8b, the information granules constructed by using all three distance measures show almost the same performance. As for Iris dataset, see Fig. 8d, while the value of c ranges from 3 to 8, the value of Q_{Acc} produced by using Manhattan distance can remain larger than the ones produced by using the other two distance measures except for $c = 7$.

V. CONCLUSIONS

This study has presented a novel granular description method based hypersphere information granules regarding three different distance measures. Compared to the existing data description methods, the proposed strategy is to construct hypersphere information granules with different geometric shapes constructed by using three different distance measures to achieve a multi-perspective description of the key geometric features of data. Three conclusions are summarized as follows: (i) using different distance measures can result in hypersphere information granules with different positions and different geometric shapes, (ii) the proposed strategy can help describe the data from multiple perspectives, (iii) the resulting hypersphere information granules have

simple architecture and reliable performance to describe the data. Some future studies would aim at further development of using the granular description method regarding different distance measures to solve classification problems.

REFERENCES

- [1] W. Pedrycz and A. Bargiela, "An optimization of allocation of information granularity in the interpretation of data structures: Toward granular fuzzy clustering," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 42, no. 3, pp. 582–590, Jun. 2012.
- [2] W. Pedrycz, G. Succi, A. Sillitti, and J. Iljazi, "Data description: A general framework of information granules," *Knowl.-Based Syst.*, vol. 80, pp. 98–108, May 2015.
- [3] W. Pedrycz, *Granular Computing: Analysis and Design of Intelligent Systems*. Boca Raton, FL, USA: CRC Press, 2016.
- [4] J. T. Yao, A. V. Vasilakos, and W. Pedrycz, "Granular computing: Perspectives and challenges," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1977–1989, Dec. 2013.
- [5] A. Bargiela and W. Pedrycz, "Toward a theory of granular computing for human-centered information processing," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 2, pp. 320–330, Apr. 2008.
- [6] W. Pedrycz, "Granular computing for data analytics: A manifesto of human-centric computing," *IEEE/CAA J. Automatica Sinica*, vol. 5, no. 6, pp. 1025–1034, Nov. 2018.
- [7] H. Fujita, A. Gaeta, V. Loia, and F. Orcioli, "Resilience analysis of critical infrastructures: A cognitive approach based on granular computing," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1835–1848, May 2019.
- [8] W. Lu, X. Chen, W. Pedrycz, X. Liu, and J. Yang, "Using interval information granules to improve forecasting in fuzzy time series," *Int. J. Approx. Reasoning*, vol. 57, pp. 1–18, Feb. 2015.
- [9] W. Pedrycz, *Knowledge-Based Clustering: From Data to Information Granules*. Hoboken, NJ, USA: Wiley, 2005.
- [10] W. Pedrycz and W. Homenda, "Building the fundamentals of granular computing: A principle of justifiable granularity," *Appl. Soft Comput.*, vol. 13, no. 10, pp. 4209–4218, Oct. 2013.
- [11] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter, "Interval analysis," in *Applied Interval Analysis*. London, U.K.: Springer, 2001, pp. 11–43.
- [12] K. Sun, L. Liu, J. Qiu, and G. Feng, "Fuzzy adaptive finite-time fault-tolerant control for strict-feedback nonlinear systems," *IEEE Trans. Fuzzy Syst.*, early access, Jan. 13, 2020, doi: 10.1109/TFUZZ.2020.2965890.
- [13] S. K., Q. Jianbin, H. R. Karimi, and Y. Fu, "Event-triggered robust fuzzy adaptive finite-time control of nonlinear systems with prescribed performance," *IEEE Trans. Fuzzy Syst.*, early access, Mar. 6, 2020, doi: 10.1109/TFUZZ.2020.2979129.
- [14] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, Oct. 1982.
- [15] W. Pedrycz, "Shadowed sets: Representing and processing fuzzy sets," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 28, no. 1, pp. 103–109, Feb. 1998.
- [16] X. Zhu, W. Pedrycz, and Z. Li, "Granular data description: Designing ellipsoidal information granules," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4475–4484, Dec. 2017.
- [17] X. Zhu, W. Pedrycz, and Z. Li, "Granular description of data: Building information granules with the aid of the principle of justifiable granularity," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Jul. 2016, pp. 969–976.
- [18] X. Zhu, W. Pedrycz, and Z. Li, "Granular representation of data: A design of families of μ -Information granules," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 4, pp. 2107–2119, Aug. 2018.
- [19] W. Lu, W. Pedrycz, J. Yang, and X. Liu, "Granular description with multi-granularity for multidimensional data: A cone-shaped fuzzy set-based method," *IEEE Trans. Fuzzy Syst.*, early access, Apr. 6, 2020, doi: 10.1109/TFUZZ.2020.2985335.
- [20] T. Ouyang, W. Pedrycz, O. F. Reyes-Galaviz, and N. J. Pizzi, "Granular description of data structures: A two-phase design," *IEEE Trans. Cybern.*, early access, Jan. 1, 2019, doi: 10.1109/TCYB.2018.2887115.
- [21] C. Fu, W. Lu, W. Pedrycz, and J. Yang, "Rule-based granular classification: A hypersphere information granule-based method," *Knowl.-Based Syst.*, vol. 194, Apr. 2020, Art. no. 105500.
- [22] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York, NY, USA: Springer, 2013.
- [23] L. Zhang, W. Lu, X. Liu, W. Pedrycz, and C. Zhong, "Fuzzy C-means clustering of incomplete data based on probabilistic information granules of missing values," *Knowl.-Based Syst.*, vol. 99, pp. 51–70, May 2016.



FANZHONG MENG received the M.S. degree in automation from Northeast Petroleum University, Daqing, China, in 1989. He joined SINOPEC Corporation, Beijing, China, in 1989, where he is currently a Senior Engineer with the Oilfield Exploration and Production Department. His current research interests include automatic control and computational intelligence.



ZHENTANG SHI received the M.E. degree from Tsinghua University, Beijing, China, in 2009. He joined the SINOPEC Dalian Research Institute of Petroleum and Petrochemicals (FRIPP), Dalian, China, in 2012, where he is currently a Senior Engineer. His current research interests include intelligent control technologies, and electricity and new energy technologies.



CHEN FU received the M.S. degree in control engineering from the Dalian University of Technology, Dalian, China, in 2013, where he is currently pursuing the Ph.D. degree with the School of Control Science and Engineering. His current research interests include granular computing, pattern recognition, and computational intelligence.



WEI LU (Member, IEEE) received the M.S. and Ph.D. degrees in control theory and control engineering from the Dalian University of Technology, Dalian, China, in 2004 and 2015, respectively. In 2004, he joined the Dalian University of Technology, where he is currently an Associate Professor with the School of Control Science and Engineering. His current research interests include computational intelligence, fuzzy modeling and granular computing, knowledge discovery and data mining, and fuzzy intelligent systems. He also serves as a frequent Reviewer for many international journals, including the *IEEE TRANSACTIONS ON FUZZY SYSTEMS*, the *IEEE TRANSACTIONS ON CYBERNETICS*, *Knowledge-Based Systems*, *Applied Soft Computing*, *Expert Systems With Applications*, the *International Journal of Approximate Reasoning*, and the *International Journal of Granular Computing* (Springer) as well as international conferences.

...